Farb- und strukturbasierte neuronale Verfahren zur Lokalisierung von Gesichtern in Real-World-Szenen*

A. Brakensiek, U.-D. Braumann, H.-J. Böhme, C. Rieck, H.-M. Gross

Technische Universität Ilmenau
Fachgebiet Neuroinformatik
D-98684 Ilmenau, Postfach 10 05 65
http://cortex.informatik.tu-ilmenau.de
anja@informatik.tu-ilmenau.de

Zusammenfassung. Für eine natürliche gestenbasierte Mensch-Maschine-Kommunikation in Real-World-Umgebungen ist die robuste Lokalisation eines potentiellen Nutzers eine elementare Voraussetzung. Dabei werden insbesondere Gesichter und Hände als besonders gestenrelevante Bildstrukturen angesehen. Der vorliegende Beitrag behandelt einen hautfarb- und einen strukturbasierten, neuronalen Lokalisationsmechanismus und stellt erste Ergebnisse zu beiden Ansätzen vor. Es wird verdeutlicht, daß erst die Kombination der beiden, sich ergänzenden Verfahren die erforderliche *Robustheit* der Nutzerlokalisation unter Real-World-Bedingungen sichert.

1 Einleitung

Die Lokalisation von gestenrelevanten Regionen (Gesichtern und Händen) eines potentiellen Nutzers ist Gegenstand verschiedener Projekte am Fachgebiet Neuroinformatik, deren gemeinsames Ziel in der Entwicklung einer neuronalen Architektur zur visuellen, gestenbasierten Interaktion eines Nutzers mit unserem Robotersystem MILVA[1] in Real-World-Umgebungen besteht ([1], vgl. auch Ansätze in [2], [3]).
Zur eigentlichen Kommunikation mit dem Nutzer dient ein aktives Zweikamerasystem, welches 7 Freiheitsgrade besitzt (für jede Kamera jeweils Pan/Tilt/Zoom, zusätzlich Pan für beide Kameras). Dieses Zweikamerasystem beobachtet aktiv die Einsatzumgebung der Roboterplattform.

* gefördert durch das Thüringer Ministerium für Wissenschaft, Forschung und Kultur (TMWFK), GESTIK-Projekt

[1] http://cortex.informatik.tu-ilmenau.de/technik.html

2 Farb- und strukturbasierte Gesichts- und Handlokalisation

2.1 Modellarchitektur

Abbildung 1 zeigt schematisch den Teil der Architektur, der die hautfarb- und gesichtsstrukturbasierte Lokalisation eines möglichen Nutzers in der Einsatzumgebung des Roboters realisiert.
Zunächst operieren beide Kameras im Weitwinkelmodus, um so einen möglichst großen Bereich der Einsatzumgebung zu erfassen. Die Bilder beider Kameras werden in jeweils einer Auflösungspyramide (Unterabtastung jeweils Faktor $1/\sqrt{2}$) in eine Multiskalenrepräsentation überführt. Auf allen Ebenen der Auflösungspyramiden operieren derzeit jeweils 2 *Cue-Module*, die sensitiv auf *Hautfarbe* und *Gesichtsstruktur* reagieren.

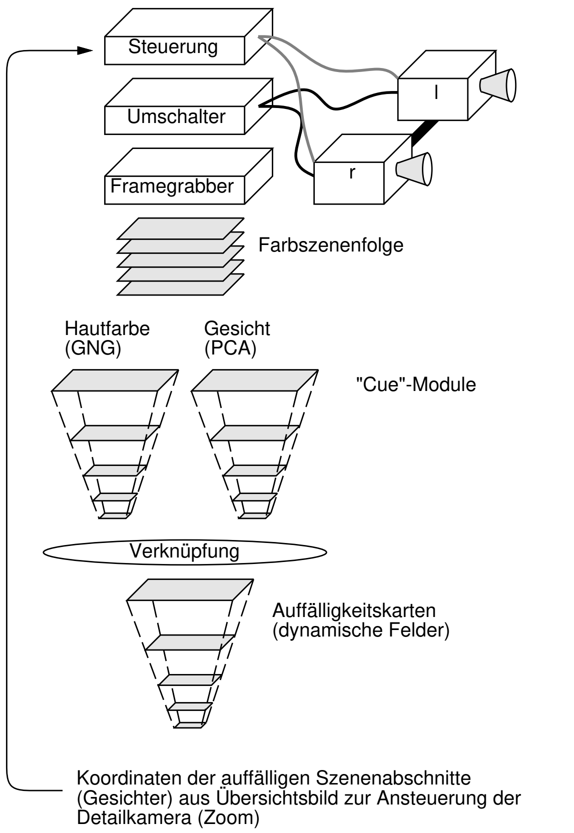

Abbildung 1. Prinzipdarstellung des Active-Vision-Systems zur Lokalisierung von Gesichtern und Händen in Real-World-Szenen: beide Kameras sind für unabhängige Übersichts- und Detailaufnahmen ausgelegt

Sobald in einem der beiden Kamerabilder ein potentieller Nutzer (Gesicht) ausgemacht wird, fungiert die entsprechende Kamera als Szenenüberblickskamera, während mit Hilfe der zweiten Kamera, die dann als Gestenkamera fungiert, die detaillierte Analyse der Gestenszene (Gesichtsverifikation, Schätzung der Gesichtsorientierung, Lokalisation der Hände, Beschreibung der Handorientierungen, Interpretation der gesamten Geste) erfolgt, worauf hier nicht näher eingegangen wird (siehe [1]). Der vorliegende Beitrag stellt die beiden oben genannten Cue-Module vor und verdeutlicht, daß erst durch deren Kombination die erforderliche *Robustheit* der Nutzerlokalisation unter Real-World-Bedingungen gesichert werden kann.

2.2 Hautfarbklassifikation

Elementarfarbraum und Farbadaptation
Für die Erstellung eines Lerndatensatzes wurden per Hand segmentierte Hautpartien von Portraitaufnahmen verwendet, die unter günstigen (relativ konstanten) Beleuchtungsverhältnissen aufgenommen wurden.

Die RGB-Farbwerte werden in einen physiologisch motivierten Elementarfarbraum transformiert (Lineartransformation, siehe [8]), der durch eine Rot-Grün-, Blau-Gelb- und Weiß-Schwarz-Achse aufgespannt wird. Die im Farbraum dargestellten Pixel eines Farbbildes ergeben eine Punktwolke, deren größte Ausdehnung bei optimaler Beleuchtung entlang der Unbuntachse verläuft. Durch eine Farbadaptation können Bilder, deren Punktwolken aufgrund nichtoptimaler Beleuchtungsverhältnisse aus der Unbuntachse ausgelenkt sind, so adaptiert werden, daß weitgehend reproduzierbare Farbwerte vorliegen (siehe [8]). Die Lage der Hautfarbpunktwolke des von uns verwendeten Datensatzes nach erfolgter Farbadaptation ist in Abbildung 2 dargestellt.

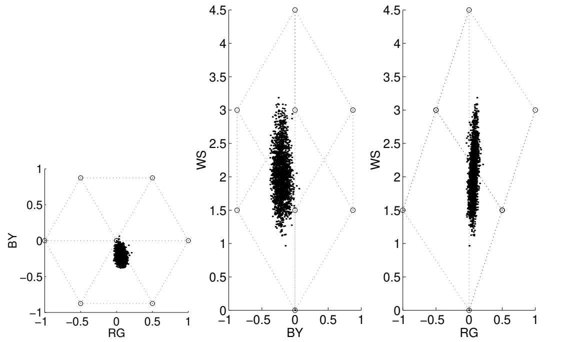

Abbildung 2. Hautfarbwolke des Trainingsdatensatzes im neurophysiologisch motivierten Elementarfarbraum: Rot-Grün/Blau-Gelb-, Blau-Gelb/Weiß-Schwarz-, Rot-Grün/Weiß-Schwarz-Farbebene;

linke Abb.: Farben der Eckpunkte im Uhrzeigersinn (von rechts): Magenta-Rot, Orange-Rot, Gelb, Grün, Cyan-Blau, Violett-Blau

Hautfarbklassifikation mit dem Growing-Neural-Gas-Netzwerk
Zur Schätzung der Wahrscheinlichkeit, mit der ein bestimmtes Pixel hautfarben ist, wird ein Growing-Neural-Gas-Netzwerk (GNG, [4]) eingesetzt, welches mit den Farbwerten der manuell segmentierten Hautpartien trainiert wird.
Das GNG im unüberwachten Modus, der hier zur Anwendung kommt, besitzt starke Ähnlichkeit mit LVQ-Verfahren und verbindet Eigenschaften der Growing-Cell-Structures mit denen des Neural-Gas (siehe auch [5]). Die Aktivität y eines GNG-Knotens i ist eine Funktion des Abstandes zwischen Eingangsvektor $\underline{x}$ und Referenzvektor $\underline{w}_i$.
Ein wesentlicher Vorteil dieses topologiebeschreibenden Netzwerkes ist die selbständige Anpassung der Netztopologie in Abhängigkeit vom lokalen Approximationsfehler. Das Netz wird mit 2 Neuronen als Minimalkonfiguration initialisiert,

der Approximationsfehler steuert das Einfügen und Löschen von Neuronen und Kanten. Kanten existieren zwischen Neuronen, deren Zuständigkeitsbereiche im Eingangssignalraum benachbart liegen. Die mittlere Länge aller Kanten eines GNG-Knotens bestimmt die Größe seines Zuständigkeitsbereichs. Für eine genaue Beschreibung des GNG sei auf [4] verwiesen.

Abbildung 3 verdeutlicht, wie sich die Knoten des GNG bzgl. ihrer Gewichtsvektoren in der Rot-Grün/Blau-Gelb-Ebene anordnen. Der Verzicht auf die Schwarz-Weiß-Ausdehnung hat den Vorteil, daß beim realen Einsatz des Netzwerks zur Segmentierung eine größere Variation der Hautfarbe zulässig ist als die, die im Trainingsdatensatz vorlag, andererseits den Nachteil einer größeren Menge von falsch klassifizierten Hintergrundregionen, die in ihrer Farbsättigung und dem Farbton der Hautfarbe entsprechen. Um jedoch sicherzustellen, daß die eventuell vorhandenen Gesichtsregionen mit großer Wahrscheinlichkeit auch als hautfarben detektiert werden (vgl. [6]), wird vorerst auf die Schwarz-Weiß-Dimension verzichtet.

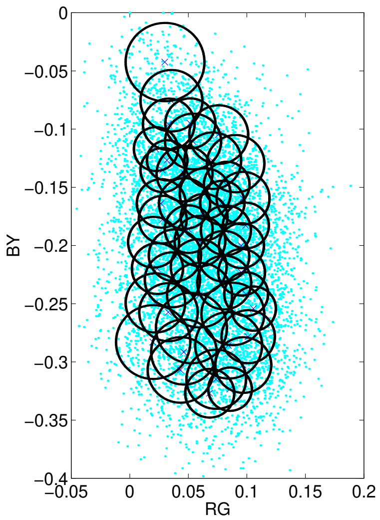

Abbildung 3. Anordnung der Zuständigkeitsbereiche der GNG-Knoten über der Hautfarbwolke

Da beim Training des GNG, welches dann als spezifisches Hautfarbmodell fungiert, die statistische Verteilung der Farbwerte (Häufigkeit des Auftretens einer Hautfarbe) mit repräsentiert wird, ergibt sich in den Gebieten häufiger Hautfarbwerte auch eine entsprechend höhere Dichte der Gewichts(referenz)vektoren der GNG-Knoten. Über die Ausgabefunktion (Aktivität) der GNG-Knoten wird implizit diese Dichte bei der Segmentierung wieder berücksichtigt, was zu hohen Hautfarbwahrscheinlichkeiten für die entsprechenden Farbwerte führt.

Eine sehr sichere Hautfarbklassifikation konnte beispielsweise in der folgenden Aufnahme (Abb. 4) durchgeführt werden. Dies ist jedoch nicht die Regel, aber bei unserem Verfahren auch nicht notwendig, da die Hautfarbe nur *einen* Beitrag zur Lokalisierung liefert.

2.3 Gesichtsstrukturdetektion

Da die Entfernung der zu lokalisierenden Person zur Kamera und somit die Gesichtsgröße nicht bekannt ist, muß die Strukturdetektion auf allen Ebenen der

Abbildung 4. Hautfarbklassifikation; Originalbild (links) und Hautklassifikationsergebnis (rechts) jeweils als Grauwertbild; die dunkelsten Pixel entsprechen den größten Hautfarbwahrscheinlichkeiten

Auflösungspyramide stattfinden (siehe Abb. 1). In unserem Szenario sei eine Person *dann* ein potentieller Nutzer, wenn ihr Gesicht frontal zum Roboter zeigt. Die Detektion von Gesichtern in Grauwertbildern erfolgt parallel zur Hautklassifikation mit Hilfe von Eigenfaces (Principal Component Analysis, PCA; vgl. [7]). Für die Berechnung der Eigenvektoren wurden die Bilder der ORL-Datenbank[2]

Abbildung 5. Mittelwert der positiven Trainingspattern (normierte Darstellung)

verwendet, von denen ein 15x15 Pixel großer Gesichtsausschnitt gewählt wurde (Abb. 5). Diese Trainingsbilder werden hinsichtlich ihres Mittelwertes und ihrer Standardabweichung normiert (vgl. [9], [10]) und mit den ersten 3 eigenwertgrößten Eigenvektoren gefaltet. Dieser Schritt der Vorverarbeitung und Merkmalsextraktion muß entsprechend auch bei der Klassifikation für jedes Fenster angewandt werden. Neben der Vorverarbeitung der Bilder spielt die Auswahl eines geeigneten Klassifikationsmaßes eine entscheidende Rolle. Als günstiges Verfahren hat sich die Klassifikation der Fitwertvektoren der Bildausschnitte mit einem Backpropagation-Netzwerk oder einem überwachten GNG erwiesen. Das Netzwerk wird mit den errechneten Fitwertvektoren der Gesichtsausschnitte und den Fitwertvektoren von ausgewählten Negativbeispielen trainiert (100 Positiv- und 100 Negativ-Beispiele). Um eine hinreichend gute Generalisierung zu erzielen, wurde ein Bootstrap-Algorithmus genutzt (siehe auch [9]), der die falsch positiv klassifizierten Bildregionen automatisch der Menge der Negativbeispiele zufügt und damit die Ausprägung scharfer Klassengrenzen fördert. Auf eine aufwendigere Vorverarbeitung der Bilder wie z.B. in [12] wurde hier verzichtet, wobei die dadurch entstehenden Unsicherheiten bei der Strukturdetektion durch die Verknüpfung mit der Hautfarberkennung wieder ausgeglichen werden. Klassifikationsergebnisse mit 3 Eigenvektoren zeigt Abb. 6. Die meisten Gesichter dieses Gruppenbildes (Ausschnitt eines Bildes von [9]) konnten lokalisiert werden, es gibt jedoch auch Fehlklassifikationen (z. B. oben rechts). Ist die eigentliche Gesichtsstruktur (Augen, Mund) nicht deutlich erkennbar oder die Auflösung zu gering, versagt die PCA.

[2] http://www.cam-orl.co.uk/face/database.html

Abbildung 6. Ergebnis der strukturbasierten Gesichtsdetektion; es wurde jeweils dort die Maske eingeblendet (invertiert), wo mit einer hohen Wahrscheinlichkeit (>0.7) ein Gesicht detektiert wurde

2.4 Parallele Nutzung beider Verfahren

Abhängig von den Umgebungsbedingungen (Beleuchtung, Szeneninhalt), die bei der Interaktion zwischen Roboter und Nutzer kaum beeinflußbar sind, bewirken die beiden beschriebenen Verfahren unterschiedlich sichere Gesichtslokalisationen. Probleme, die sich bei der Hautfarbklassifikation in Real-World-Szenen ergeben, werden in dem folgenden Beispiel deutlich (Abb. 7, obere Reihe), bei dem die strukturbasierte Lokalisation jedoch eindeutig ist. Eine optimale Lokalisation durch beide vorgestellte Verfahren zeigt die untere Reihe in Abb. 7. Erst die parallele Nutzung der beiden, sich gegenseitig ergänzenden Verfahren sichert die notwendige Robustheit der Nutzerlokalisation auch unter hochgradig variablen Umgebungsbedingungen. Durch den Einsatz der parallel operierenden Cue-Module wird das Gesamtsystem damit weniger abhängig vom Vorhandensein *einer bestimmten* Informationsqualität im Bild.

3 Ausblick

Um die Robustheit der Nutzerlokalisation weiter zu erhöhen, werden derzeit zwei zusätzliche Cue-Module zur Detektion von Bewegung und zur Detektion einer Kopf-Schulter-Partie jeweils über dem Grauwertbild implementiert. Die Ergebnisse aller 4 Cue-Module werden dann in topographisch organisierte Felder dynamischer Neuronen eingekoppelt ([11], siehe auch Abbildung 1), die die Aufmerksamkeit der Gestenkamera auf die Regionen richten sollen, in denen mit großer Wahrscheinlichkeit ein Gesicht liegt. Die dazu notwendige Entfernungschätzung liefert dabei die Strukturinformation (Eigenfaces bzw. Kopf-Schulter-Partie).
Da eine optimale Farbadaptation nach [8] ein möglichst breites Farbspektrum des Bildinhalts voraussetzt, was gerade unter den vorliegenden indoor-Bedingungen

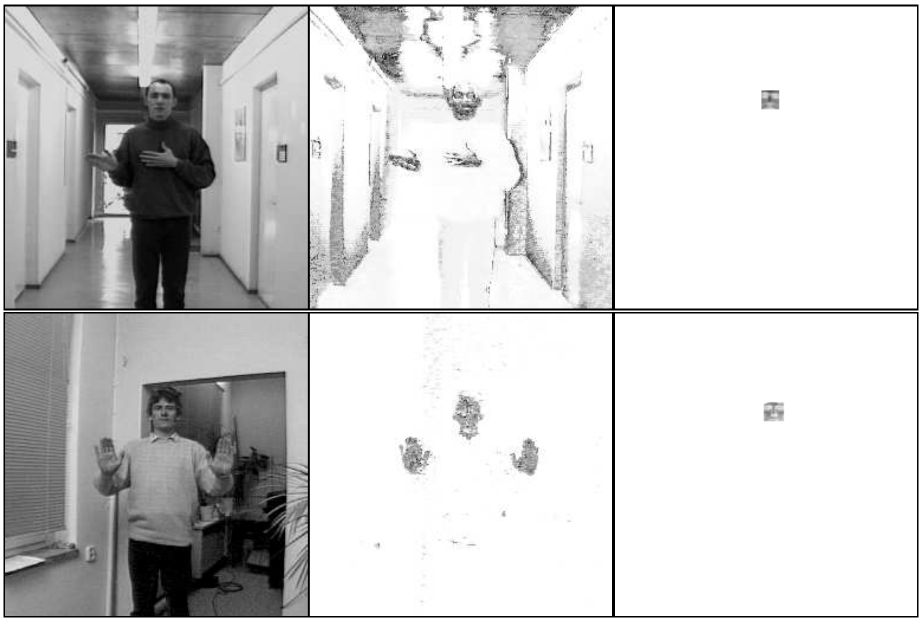

Abbildung 7. Haut- (mitte) und Strukturhypothesenbilder (rechts) farbadaptierter Bilder (links), die mit einem GNG (49 Neurone) bzw. einer PCA mit 3 Eigenvektoren bei unterschiedlicher Beleuchtung (Leuchtstoffröhre (oben), Halogenscheinwerfer (unten)) und unterschiedlichem Hintergrund erzielt wurden

oft nur unzureichend zutrifft, soll die Farbadaptation so angepaßt werden, daß die Punktwolken in eine Richtung adaptiert werden, die sich aus der Differenz der Lage des gelernten Hautfarbmodells und der aktuellen Hautfarbwolke eines einmal detektierten und verifizierten Gesichtes ergibt, was einer Farbadaption mit dem gelernten Hautfarbmodell als impliziter Referenz entspricht. Ziel dieses Verfahrens ist es, die Hände des ermittelten Nutzers *sicherer* erkennen zu können, was für die sich anschließende Gestenerkennung notwendig ist.
Abbildung 8 verdeutlicht, daß bei farblich unausgewogenem Bildinhalt (hier: gelb-grün, Abb. 8 links) die Punktwolke der Farbwerte nach der Farbadaption nach [8] zwar entlang der Unbuntachse des Elementarfarbraumes, die Hautfarbwolke jedoch leicht in Blaurichtung verläuft (Abb. 8 rechts). Die Hautfarbe würde in diesem Fall nur unzureichend klassifiziert werden (vgl. Referenzhautwolke in Abb. 2 bzw. GNG in Abb. 3). Nutzt man dagegen das gelernte Hautfarbmodell und adaptiert die Farbwerte so, daß die Hautfarbwerte des verifizierten Gesichtes diesem Modell weitestgehend entsprechen, so ergibt sich ein Bild, welches die "realen" Farbverhältnisse besser widerspiegelt.

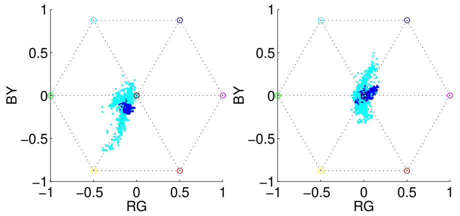

Abbildung 8. Farbwolke (Hintergrund: hell, Hautfarbe: dunkel) des Originalbildes (links) und des nach [8] farbadaptierten Bildes (rechts) in der Rot-Grün / Blau-Gelb-Farbebene

Danksagung
Die Autoren danken Markus Krabbes, Rolf Nestler und Thomas Kleemann für konstruktive Diskussionen und Hinweise sowie die technische Unterstützung.

Literatur

1. Böhme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., Gross, H.-M.: Neural architecture for gesture-based human-machine-interaction. eingereicht bei: Bielefeld Gesture Workshop, 1997
2. Crowley, J. L., Coutaz, J.: Vision for Man Machine Interaction. EHCI, Grand Targhee, 1995
3. Darrell, T., Basu, S., Wren, C., Pentland, A.: Perceptually-driven Avatars and Interfaces: active methods for direct control. M.I.T. Media Lab Perceptual Computing Section Technical Report No.416, 1997
4. Fritzke, B.: A Growing Neural Gas Network Learns Topologies. In: Proc. of NIPS 7, pp. 625-632, 1995
5. Hamker, F., Heinke, D.: Implementation and Comparison of Growing Neural Gas, Growing Cell Structures and Fuzzy Artmap. Schriftenreihe des Fachgebietes Neuroinformatik der TU Ilmenau, Report Nr.1/97, 1997
6. Littmann, E., Ritter, H.: Neural and Statistical Methods for Adaptive Color Segmentation - A Comparison. In: Mustererkennung 1995, S. 84-93, 1995
7. Moghaddam, B., Pentland, A.: Maximum Likelihood Detection of Faces and Hands. In: Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, pp. 122-128, 1995
8. Pomierski, T., Gross, H.-M.: Biological Neural Architecture for Chromatic Adaptation Resulting in Constant Color Sensations. In: Proc. of ICNN'96, IEEE, pp. 734-739, 1996
9. Rowley, H. A., Baluja, S. Kanade, T.: Human Face Detection in Visual Scenes. Technical Report CMU-CS-95-158R, 1995
10. Schiele, B., Waibel, A.: Gaze Tracking Based on Face-Color. In: Int. Workshop on Automatic Face- and Gesture-Recognition, Zürich, pp. 344-349, 1995
11. Stephan, V., Groß, H.-M.: Formerhaltende sequentielle visuelle Aufmerksamkeit in columnar organisierten neuronalen Feldern. DAGM, 1997
12. Sung, K.-K., Poggio, T.: Example-based Learning for View-based Human Face Detection. Proceedings Image Understanding Workshop, 1994

This article was processed using the LaTeX macro package with LLNCS style